Equipercentile Equating with Equal Interval Scores

Brad Hanson
February 10, 1993 (revised 5/2/95)

Let X and Y be discrete random variables representing the distribution of scores on two forms of a test (labeled Form X and Form Y, respectively) in the some population. The possible values taken on by X are $s_x(i), i = 0, \ldots, K_x$, where there are $K_x + 1$ possible scores for Form X. The possible values taken on by Y are $s_y(j),\ j = 0, \ldots, K_y$, where there are $K_y + 1$ possible scores for Form Y. It is assumed that $s_x(i) - s_x(i-1) = d_x > 0,\ i = 1, \ldots, K_x$ and $s_y(j) - s_y(j-1) = d_y > 0,\ j = 1, \ldots, K_y$.

The goal of equipercentile equating is to find a function e such that $G[e(x)] = F(x)$ for all x and $F[e^{-1}(y)] = G(y)$ for all y, where F and G are the cumulative distributions for X and Y, respectively. If X and Y were continuous then the equipercentile conversion of scores on Form X to scores on Form Y would be given by $e(x) = G^{-1}F(x)$ and the equipercentile conversion of scores on Form Y to Form X would be given by $e^{-1}(y) = F^{-1}G(y)$. Since X and Y are discrete random variables, F^{-1} and G^{-1} are not defined and consequently the equipercentile equating functions are not defined.

One way to define an equipercentile equating function for discrete test scores is to use continuous approximations of X and Y in place of the discrete distributions. The equipercentile equating function is defined in terms of the continuous approximations and applied to the discrete test scores.

One possibility for obtaining continuous approximations to the discrete score distributions is to use kernel estimators (Holland and Thayer, 1989). Kernel estimators are used to provide estimates of continuous distributions from a sample containing discrete data. Kernel estimators are typically used in cases where the underlying distribution is known to be continuous and an estimate of the continuous distribution is desired based on a sample of discrete data values. In this paper the kernel estimators are used to create an artificial continuization of a distribution which is discrete so that the inverse of the distribution function is defined.

A kernel continuization of the discrete random variable X is given by

$$f_k(x) = \frac{1}{h}\sum_{i=0}^{K_x} \kappa\left(\frac{x - s_x(i)}{h}\right) f[s_x(i)]\,, \tag{1}$$

where $f[s_x(i)] = \Pr[X = s_x(i)]$ (it is assumed $f[s_x(i)] > 0$ for all i), h is a parameter determining the degree to which the discrete density is spread out, and $\kappa(x)$ is a continuous (usually symmetric) density function on the interval $[-a, a]$. The kernel continuization $f_k[s_x(i)]$ is greater than 0 in the interval $[s_x(0) - ah, s_x(K_x) + ah]$ and is equal to 0 outside this interval. The function $f_k(x)$ is a density function on the interval $[s_x(0) - ah, s_x(K) + ah]$ since

$$\begin{aligned}
\int_{s_x(0)-ah}^{s_x(K_x)+ah} f_k(x)dx &= \int_{s_x(0)-ah}^{s_x(K_x)+ah} \frac{1}{h}\sum_{i=0}^{K_x} \kappa\left(\frac{x - s_x(i)}{h}\right) f[s_x(i)]dx \\
&= \sum_{i=0}^{K_x} \int_{s_x(0)-ah}^{s_x(K_x)+ah} \frac{1}{h}\kappa\left(\frac{x - s_x(i)}{h}\right) dx f[s_x(i)] \\
&= \sum_{i=0}^{K_x} \int_{[s_x(0)-ah-s_x(i)]/h}^{[s_x(K_x)+ah-s_x(i)]/h} \kappa(u)du f[s_x(i)]\,,
\end{aligned}$$

where $u = [x - s_x(i)]/h$. For all i, $[s_x(0) - ah - s_x(i)]/h \leq -a$ and $[s_x(K) + ah - s_x(i)]/h \geq a$. Consequently,

$$\sum_{i=0}^{K_x} \int_{[s_x(0)-ah-s_x(i)]/h}^{[s_x(K_x)+ah-s_x(i)]/h} \kappa(u) du f[s_x(i)] = \sum_{i=0}^{K_x} f[s_x(i)]$$
$$= 1\,.$$

Consider the case where $\kappa(x) = 1/d_x$ for x in the interval $[-.5d_x, .5d_x]$ (κ is a uniform density). In this case the cumulative distribution function of the kernel continuization of X ($F_k(z)$) is given by

$$\begin{aligned} F_k(z) &= \int_{s_x(0)-.5d_xh}^{z} f_k(x) dx \\ &= \int_{s_x(0)-.5d_xh}^{z} \frac{1}{h} \sum_{i=0}^{K_x} \kappa\left(\frac{x - s_x(i)}{h}\right) f[s_x(i)] dx \\ &= \sum_{i=0}^{K_x} \int_{s_x(0)-.5d_xh}^{z} \frac{1}{h} \kappa\left(\frac{x - s_x(i)}{h}\right) dx f[s_x(i)] \\ &= \sum_{i=0}^{K_x} \int_{[s_x(0)-.5d_xh-s_x(i)]/h}^{[z-s_x(i)]/h} \kappa(u) du f[s_x(i)]\,, \end{aligned} \qquad (2)$$

where $u = [x - s_x(i)]/h$ and it is assumed that $h \geq 1$. The lower limit of the integral in Equation 2 will be less than or equal to $-.5d_x$ for all i. For $s_x(i) \leq z - .5d_xh$ the upper limit of the integral in Equation 2 is greater than or equal to $.5d_x$. Consequently, for $s_x(i) \leq z - .5d_xh$ the integral in Equation 2 can be written as

$$\int_{-.5d_x}^{.5d_x} \frac{1}{d_x} du = 1\,.$$

For $s_x(i) \geq z + .5d_xh$ the upper limit of the integral in Equation 2 is less than or equal to $-.5d_x$ so that the integral in Equation 2 is equal to zero. For $z - .5d_xh < s_x(i) < z + .5d_xh$ the integral in Equation 2 can be written as

$$\int_{-.5d_x}^{[z-s_x(i)]/h} \frac{1}{d_x} du = \left(\frac{z - s_x(i)}{d_xh} + .5\right)\,.$$

Let i_l^* be the smallest integer such that $s_x(i_l^*) > z - .5d_xh$ and let i_u^* be the largest integer such that $s_x(i_u^*) < z + .5h$. Equation 2 can then be written as

$$F_k(z) = \sum_{i=0}^{i_l^*-1} f[s_x(i)] + \sum_{i=i_l^*}^{i_u^*} \left(\frac{z - s_x(i)}{d_xh} + .5\right) f[s_x(i)]\,. \qquad (3)$$

If $h = 1$ then $i_l^* = i_u^* = i^*$ and Equation 3 can be written as

$$\begin{aligned} F_k(z) &= \sum_{i=0}^{i^*-1} f[s_x(i)] + \left(\frac{z - s_x(i^*)}{d_x} + .5\right) f[s_x(i^*)] \\ &= F[s_x(i^* - 1)] + \left(\frac{z - s_x(i^*)}{d_x} + .5\right) f[s_x(i^*)]\,. \end{aligned} \qquad (4)$$

Equation 4 multiplied by 100 is the percentile rank function. This function gives percentile ranks as defined in many elementary statistics and measurement texts (e.g. Blommers and Forsyth, 1977). Consequently, the traditional definition of percentile rank is seen as equivalent to the percentiles of a continuization of the discrete score distribution using a uniform kernel continuization with $h = 1$ (Holland and Thayer, 1989).

The equipercentile equating function for converting scores on Form X to scores on Form Y using the kernel continuization given in Equation 4 is $G_k^{-1} F_k(x)$, where $G_k(y)$ is the analog of Equation 4 for Y given by

$$G_k(z) = G[s_y(j^* - 1)] + \left(\frac{z - s_y(j^*)}{d_y} + .5 \right) g[s_y(j^*)] \,. \tag{5}$$

where $g[s_y(j)] = \Pr[Y = s_y(j)]$ and j^* is the smallest integer such that $s_y(j^*) > z - .5d_y$.

For a value p in the interval [0, 1] the value z such that $G_k(z) = p$ is given by

$$G_k^{-1}(p) = d_y \left(\frac{p - G[s_y(j^* - 1)]}{g[s_y(j^*)]} \right) - .5d_y + s_y(j^*) \,, \tag{6}$$

where j^* is the smallest integer such that $p < G[s_y(j^*)]$. Using Equations 4 and 6 the equipercentile equating function for converting scores on Form X to scores on Form Y is

$$G_k^{-1} F_k[s_x(i)] = d_y \left(\frac{p^*(i) - G[s_y(j^* - 1)]}{g[s_y(j^*)]} \right) - .5d_y + s_y(j^*) \tag{7}$$

where

$$p^*(i) = F[s_x(i-1)] + .5 f[s_x(i)] \,,$$

and j^* is the smallest integer such that $p^*(i) < G[s_y(j^*)]$.

In applications the discrete score density $f[s_x(i)]$ is not known (the following discussion of $f[s_x(i)]$ also applies to $g[s_y(j)]$). Instead, an estimate of $f[s_x(i)]$ would be used. One possibility is to use the observed score probabilities as estimates of $f[s_x(i)]$. A possible problem is that if any of observed probabilities are zero then $\hat{f}_k(x) = 0$ for some x in the interval $[s_x(0) - .5d_x, s_x(K) + .5d_x]$. This produces the same problem as the continuization was supposed to solve — the cumulative distribution function has the same value for multiple values of x and consequently the inverse of the cumulative distribution function is undefined.

One possible way to avoid this problem is to use an estimate of $f[s_x(i)]$ that is greater than zero for all i, even if there are some scores that are not obtained by any examinee in the sample. Such an estimate may be obtained by fitting a model to the data, or by mixing the observed distribution with a uniform distribution, using a large mixing weight for the observed distribution and a corresponding small mixing weight for the uniform distribution.

Another possibility is to pick one of the possible values for the inverse of the cumulative distribution function when there are multiple values of x which produce the same value of $\hat{F}_k(x)$. For example, if all values in an interval $[x_l, x_u]$ produce the same value of $\hat{F}_k(x)$ then the midpoint of the interval $(.5x_l + .5x_u)$ might be used as the inverse of $\hat{F}_k(x)$ for x in the interval $[x_l, x_u]$.

The use of a kernel function $[\kappa(x)]$ that was non-zero on a larger interval than $[-.5d_x, .5d_x]$ could also be used to prevent zero values of $\hat{f}_k(x)$. For example, Holland and Thayer (1989) suggest using a normal density function for $\kappa(x)$. This would result in $\hat{f}_k(x) > 0$ for all real x.

A value of h greater than 1 could be used with the uniform kernel function to eliminate zero values of $\hat{f}_k(x)$. A variable kernel (where there are different values of h corresponding to each score) could be used to spread the density more for scores close to scores with zero density.

References

Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education (2nd ed.)*. Boston, MA: Houghton Mifflin.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions*. ETS Research Report No. 89-7. Princeton NJ: Educational Testing Service.